Canonical Correlation Analysis for Longitudinal Data

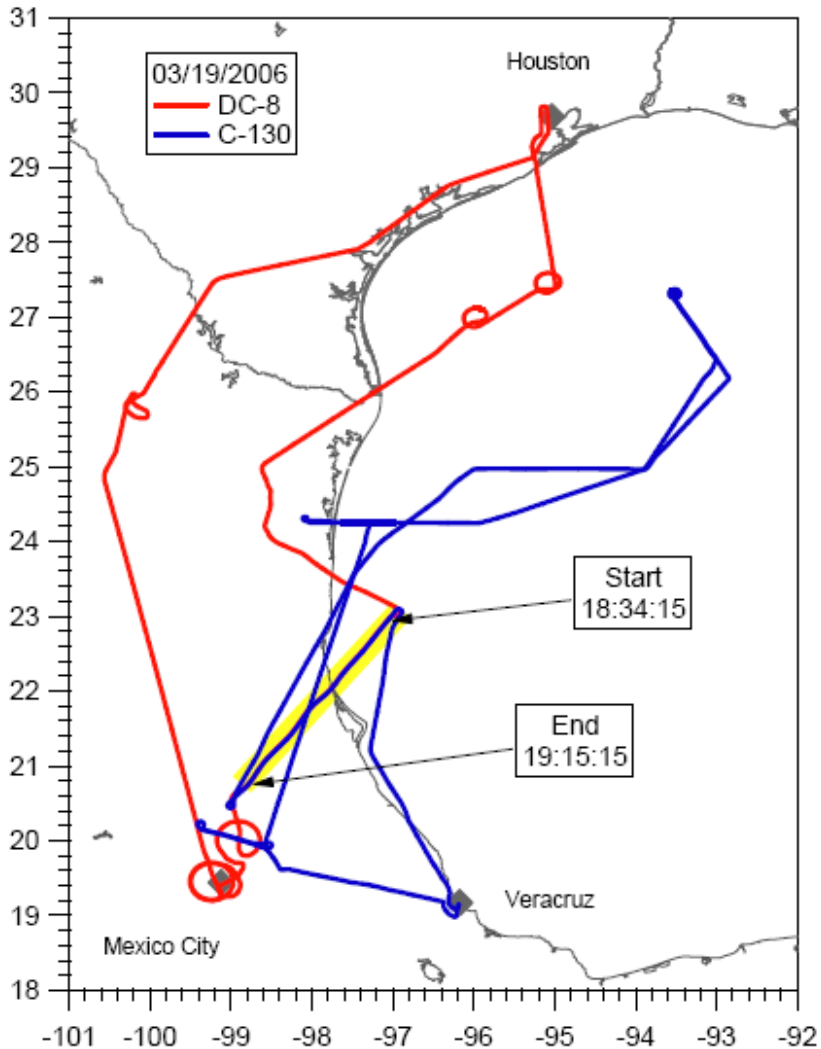
Raymond McCollum
Advisor Dayanand Naik

May 5th, 2010

INTEX

- The Intercontinental Chemical Transport Experiment (INTEX)
- "INTEX (http://cloud1.arc.nasa.gov) is a two phase experiment that aims to understand the transport and transformation of gases and aerosols on transcontinental/intercontinental scales and assess their impact on air quality and climate."
- The experiment was performed in the spring of 2006.
- The purpose of the project was to "Quantify the outflow and evolution of gases and aerosols from the Mexico City Megaplex".

Analysis Air Tracks*



INTEX*

- Multiple air frames will measure air and pollutants along the Mexican Coast.
- NASA DC-8 flown out o Houston, Texas
- NSF/NCAR C-130 from Tampico, Mexico
- Air frames will travel in close proximity.
- Data from multiple gasses will be recorded for each plane and compared in an effort to calibrate the instrumentation.

Canonical Correlation

Canonical correlation analysis (CCA) is used to identify and characterize the relationship between two sets of random vectors. Let Σ_x be the variance-covariance matrix of the $px1$ vector X, and let Σ_y be the variance covariance matrix of $qx1$ vector Y. Let the multivariate vectors X and Y have the covariance matrix Σ_{xy}.

$$\begin{bmatrix} \Sigma_y & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_x \end{bmatrix} \tag{1}$$

Classical Canonical Correlation*

The objective of canonical correlation analysis is to create a relationship between the X variables and the Y variables. CCA attempts to find a $qx1$ vector "a" and a $px1$ vector "b" to help define the correlation between the two data sets. The vectors a and b are chosen to maximize the correlation between $a'Y$ and $b'X$.

Repeated Canonical Correlation*

CCA was generalized to more than two sets of variables by Kettenring (1971)and other generalizations can be found in the literature. Let X and Y be repeated over time. Let $\mathbf{x}_i$ and $\mathbf{y}_i$ be vectors observed at the i^{th} time period. Represent the time periods by,

$$Y = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_t') \text{ and } X = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_t'). \quad (2)$$

Repeated Canonical Correlation*

For a CCA the with no additional time requirements, the covariance matrix is,

$$\begin{pmatrix} \Sigma_y & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_x \end{pmatrix}. \tag{3}$$

Repeated Canonical Correlation*

The total number of parameters that require estimation becomes computationally intensive.

- The Σ_y matrix has $q(q+1)/2$ unique parameters.
- The Σ_x matrix has $p(p+1)/2$ unique parameters.
- The Σ_{xy} matrix has pq unique parameters for the cross correlations.

There are a total of $p(p+1)/2 + q(q+1)/2 + pq$ unique parameters that must be estimated. These values correspond to one set of parameters recorded at one time period. When sets of variables are recorded over time, the number of parameters required increases quickly.

Repeated Canonical Correlation*

For t time periods the matrix becomes,

$$\begin{pmatrix} \Sigma_{y1y1} & \cdots & \Sigma_{y1yt} & \Sigma_{y1x1} & \cdots & \Sigma_{y1xt} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \Sigma_{yty1} & \cdots & \Sigma_{ytyt} & \Sigma_{ytx1} & \cdots & \Sigma_{ytxt} \\ \Sigma_{x1y1} & \cdots & \Sigma_{x1yt} & \Sigma_{x1x1} & \cdots & \Sigma_{x1xt} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \Sigma_{xty1} & \cdots & \Sigma_{xtyt} & \Sigma_{xtx1} & \cdots & \Sigma_{xtxt} \end{pmatrix} \tag{4}$$

This has a total of $\frac{t(p+q)(t(p+q)+1)}{2}$ parameters.

Repeated Canonical Correlation*

The Kronecker product covariance structure can be used to reduce the number of parameters. The variance covariance matrix of Y and X can be represented by the matrix below.

$$\begin{pmatrix} \Psi_y \otimes \Sigma_y & \Psi_{yx} \otimes \Sigma_{yx} \\ \Psi_{xy} \otimes \Sigma_{xy} & \Psi_x \otimes \Sigma_x \end{pmatrix} \tag{5}$$

This matrix has considerably less parameters, namely

$$\frac{q(q+1) + p(p+1) + 2pq + 3t(t+1)}{2}. \tag{6}$$

Existing Solution*

Suppose $z_1, ..., z_N$ is a random sample of size N from the above multivariate normal distribution in (**??**). SNV (2008) obtained the maximum likelihood estimates of Σ and Ψ as follows. The MLE of an unrestricted positive definite matrix Σ is given as

$$\hat{\Sigma} = \frac{\sum_{i=1}^{N} z_{ic} \hat{\Psi}^{-1} z_{ic}'}{tN} \tag{7}$$

and similarly the MLE of an unrestricted matrix Ψ, except for the restriction that $\psi_{tt} = 1$, is given by

$$\hat{\Psi} = \frac{\sum_{i=1}^{N} z_{ic}' \hat{\Sigma}^{-1} z_{ic}}{pN} \tag{8}$$

and $\hat{\psi}_{tt} = 1$.

Existing Solution*

Here

$$z_i = \begin{pmatrix} z_{i11} & z_{i12} & \cdots & z_{i1t} \\ z_{i21} & z_{i22} & \cdots & z_{i2t} \\ \vdots & \vdots & \ddots & \vdots \\ z_{ip1} & z_{ip2} & \cdots & z_{ipt} \end{pmatrix},$$

$$z_{ic} = z_i - \bar{z}, \tag{9}$$

where,

$$\bar{z} = \begin{pmatrix} \bar{z}_{11} & \bar{z}_{12} & \cdots & \bar{z}_{1t} \\ \bar{z}_{21} & \bar{z}_{22} & \cdots & \bar{z}_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{z}_{p1} & \bar{z}_{p2} & \cdots & \bar{z}_{pt} \end{pmatrix}.$$

Existing Solution*

$$\hat{\Sigma}^{-1} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1p} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2p} \\ \vdots & \ddots & & \vdots \\ \alpha_{p1} & \alpha_{p2} & \dots & \alpha_{pp} \end{pmatrix}$$

The solution to the Ψ multiplies the z_{ic} matrix by the Σ inverse estimate,

$$\hat{\psi} = \frac{\sum_{i=1}^{N} z_{ic}' \hat{\Sigma}^{-1} z_{ic}}{Np},$$

Cross Correlation*

$$D = \begin{pmatrix} \Psi_y \otimes \Sigma_y & \Psi_{yx} \otimes \Sigma_{yx} \\ \Psi_{xy} \otimes \Sigma_{xy} & \Psi_x \otimes \Sigma_x \end{pmatrix}. \tag{10}$$

Cross Correlation*

Estimates are,

$$\hat{\Psi}_{xy} = \frac{\sum_{i=1}^{N} X_{ic}' \hat{\Sigma}_{xy}^{+} Y_{ic}}{Rank(\Sigma_{xy})N} \tag{11}$$

$$\hat{\Sigma}_{xy} = \frac{\sum_{i=1}^{N} Y_{ic}' \hat{\Psi}_{xy}^{-1} X_{ic}}{Rank(\Psi_{xy})N}. \tag{12}$$

Partitioning Σ^*

Partitioning $\mathbf{\Sigma}$ will give

$$\mathbf{\Sigma} = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}.$$

This can be broken up to get

$$\mathbf{D} = \begin{bmatrix} \Psi_{yy} \otimes \Sigma_{yy} & \Psi_{yx} \otimes \Sigma_{yx} \\ \Psi_{xy} \otimes \Sigma_{xy} & \Psi_{xx} \otimes \Sigma_{xx} \end{bmatrix}.$$

Transformation*

Transform the data to make the diagonals of the covariance matrix the identity matrix.

Let

$$\begin{bmatrix} (\hat{\Psi}_y \otimes \hat{\Sigma}_y)^{-1/2} & 0 \\ 0' & (\hat{\Psi}_x \otimes \hat{\Sigma}_x)^{-1/2} \end{bmatrix} \begin{bmatrix} Y \\ X \end{bmatrix} \tag{13}$$

Transformation*

The normal distribution was used to approximate the asymptotic distribution. Similar to Tan (1973) work.

$$\underset{\sim}{\overset{approximately}{}} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}\begin{pmatrix} I & D \\ D' & I \end{pmatrix}\right] \tag{14}$$

$$D = (\hat{\Psi}_x \otimes \hat{\Sigma}_x)^{-1/2} C (\hat{\Psi}_y \otimes \hat{\Sigma}_y)^{-1/2} \tag{15}$$

Transformation*

$$\begin{bmatrix} I & D \\ D' & I \end{bmatrix}^{-1/2} \begin{bmatrix} (\hat{\Psi}_y \otimes \hat{\Sigma}_y)^{-1/2} & 0 \\ 0' & (\hat{\Psi}_x \otimes \hat{\Sigma}_x)^{-1/2} \end{bmatrix} \begin{bmatrix} Y \\ X \end{bmatrix} \tag{16}$$

$$\underset{\sim}{\overset{approximately}{}} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}\right] \tag{17}$$

$$\begin{bmatrix} 0 & D \\ D' & 0 \end{bmatrix} \begin{bmatrix} I & D \\ D' & I \end{bmatrix}^{-1/2} \begin{bmatrix} (\hat{\Psi}_y \otimes \hat{\Sigma}_y)^{-1/2} & 0 \\ 0' & (\hat{\Psi}_x \otimes \hat{\Sigma}_x)^{-1/2} \end{bmatrix} \tag{18}$$

$$* \begin{bmatrix} Y \\ X \end{bmatrix} \underset{\sim}{\overset{approximately}{}} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} DD' & 0 \\ 0 & D'D \end{pmatrix}\right] \tag{19}$$

Unrestricted Covariance Matrix Ψ*

The solution gives,

$$\widehat{D_{\Psi_{xy}} D'_{\Psi_{xy}}} \otimes \widehat{D_{\Sigma_{xy}} D'_{\Sigma_{xy}}}. \tag{20}$$

Spectral decomposition gives,

$$\widehat{D_{\Psi_{xy}} D'_{\Psi_{xy}}} = U \Delta^2 U' \Rightarrow \tag{21}$$

For $D_{\Psi_{xy}}$ positive definite we have the unique matrix (Harville (1997)).

$$\hat{D}_{\Psi_{xy}} = U \Delta U' \tag{22}$$

Unrestricted Covariance Matrix Ψ*

$$C_{Base_{\Psi_{xy}}} = A_{\Psi_y}^{1/2} \hat{D}_{\Psi_{xy}} B_{\Psi_y}^{1/2} \tag{23}$$

Note that $C_{Base_{\Psi_{xy}}}$ can be a correlation matrix,

$$\hat{C}_{\Psi_{xy}} = Diag(\hat{C}_{Base_{\Psi_{xy}}})^{-1/2} C_{Base_{\Psi_{xy}}} Diag(\hat{C}_{Base_{\Psi_{xy}}})^{-1/2} \tag{24}$$

or $C_{Base_{\Psi_{xy}}}$ can use the AR(1) structure as noted earlier.
At this point we have a complete estimate for $\hat{C}_{\Psi_{xy}}$ to use to estimate $\hat{\Sigma_{xy}}$

$$\hat{\Sigma_{xy}} = \frac{\sum_{i=1}^{N} X_{ic} \hat{C}_{\Psi_{xy}}^{-1} Y_{ic}'}{NRank(C)} \tag{25}$$

Estimation Results*

Sample Size 500	Simulations 1000			
Parameter	Θ	$\hat{\Theta}$	$(E(\Theta - \hat{\Theta})^2)$	$\lvert(\hat{\Theta} - \Theta)\rvert$
$\Sigma_y(11)$	5.50000	5.49091	0.19355	0.00909
$\Sigma_y(12)$	3.17543	3.16755	0.16785	0.00788
$\Sigma_y(13)$	6.50000	6.48315	0.23614	0.01685
$\Sigma_y(14)$	-0.50000	-0.50270	0.15944	0.00270
$\Sigma_y(21)$	3.17543	3.16691	0.19235	0.00852
$\Sigma_y(22)$	7.50000	7.48407	0.27176	0.01593
$\Sigma_y(23)$	0.35355	0.34927	0.14595	0.00428
$\Sigma_y(31)$	2.24537	2.23774	0.17092	0.00762
$\Sigma_y(32)$	-0.35355	-0.35161	0.17407	0.00195
$\Sigma_y(33)$	6.25000	6.23379	0.23151	0.01621
ρ_y	0.20000	0.19939	0.01497	0.00061
$\Sigma_x(11)$	4.55000	4.53781	0.18379	0.01219
$\Sigma_x(12)$	0.44907	0.44612	0.09253	0.00296
$\Sigma_x(21)$	2.70000	2.69447	0.10096	0.00553
ρ_x	0.40000	0.39907	0.01952	0.00093

Estimation Results*

Sample Size 500	Simulations 1000			
Parameter	Θ	$\hat{\Theta}$	$(E(\Theta - \hat{\Theta})^2)$	$\lvert(\hat{\Theta} - \Theta)\rvert$
$\Sigma_{xy}(11)$	1.00416	1.00466	0.13027	0.00050
$\Sigma_{xy}(12)$	1.56525	1.56436	0.10695	0.00088
$\Sigma_{xy}(13)$	1.73925	1.73518	0.15310	0.00408
$\Sigma_{xy}(14)$	1.42009	1.41771	0.11413	0.00238
$\Sigma_{xy}(21)$	-0.27386	-0.27980	0.14816	0.00594
$\Sigma_{xy}(22)$	-0.22361	-0.22382	0.11304	0.00022
$\Sigma_{xy}(23)$	0.19365	0.18905	0.13287	0.00460
$\Sigma_{xy}(24)$	0.15811	0.15158	0.10581	0.00654
ρ_{xy}	0.30000	0.30217	0.04504	0.00217
$\Sigma 1_{xy}(11)$	1.00416	1.18419	0.45267	0.18003
$\Sigma 1_{xy}(12)$	1.56525	1.82970	0.61523	0.26446
$\Sigma 1_{xy}(13)$	1.73925	2.02502	0.67560	0.28576
$\Sigma 1_{xy}(14)$	1.42009	1.66430	0.57283	0.24420
$\Sigma 1_{xy}(21)$	-0.27386	-0.33653	0.21721	0.06267
$\Sigma 1_{xy}(22)$	-0.22361	-0.26029	0.16079	0.03669
$\Sigma 1_{xy}(23)$	0.19365	0.21089	0.16858	0.01724
$\Sigma 1_{xy}(24)$	0.15811	0.18202	0.14543	0.02391
$\rho_1 xy$	0.30000	0.34486	0.10577	0.04486

Estimation Results*

Sample Size 50	Simulations 1000			
Parameter	Θ	$\hat{\Theta}$	$(E(\Theta - \hat{\Theta})^2)$	$\lvert(\hat{\Theta} - \Theta)\rvert$
$\Sigma_y(11)$	5.50000	5.38572	0.63838	0.11428
$\Sigma_y(12)$	3.17543	3.12812	0.53214	0.04730
$\Sigma_y(13)$	6.50000	6.37219	0.75108	0.12781
$\Sigma_y(14)$	-0.50000	-0.48519	0.52173	0.01481
$\Sigma_y(21)$	3.17543	3.10831	0.63379	0.06712
$\Sigma_y(22)$	7.50000	7.35267	0.87776	0.14734
$\Sigma_y(23)$	0.35355	0.35696	0.46468	0.00340
$\Sigma_y(31)$	2.24537	2.20256	0.55177	0.04281
$\Sigma_y(32)$	-0.35355	-0.35449	0.55607	0.00094
$\Sigma_y(33)$	6.25000	6.12585	0.74061	0.12415
ρ_y	0.20000	0.19905	0.04838	0.00095
$\Sigma_x(11)$	4.55000	4.43166	0.56430	0.11834
$\Sigma_x(12)$	0.44907	0.44469	0.29620	0.00438
$\Sigma_x(21)$	2.70000	2.65309	0.32754	0.04691
ρ_x	0.40000	0.39930	0.06129	0.00070

Estimation Results*

Sample Size 50	Simulations 1000	Converge		
Parameter	Θ	$\hat{\Theta}$	$(E(\Theta - \hat{\Theta})^2)$	$\lvert(\hat{\Theta} - \Theta)\rvert$
$\Sigma_{xy}(11)$	1.00416	0.99012	0.40499	0.01404
$\Sigma_{xy}(12)$	1.56525	1.55870	0.34521	0.00655
$\Sigma_{xy}(13)$	1.73925	1.70115	0.46706	0.03810
$\Sigma_{xy}(14)$	1.42009	1.41299	0.36314	0.00710
$\Sigma_{xy}(21)$	-0.27386	-0.27898	0.48118	0.00512
$\Sigma_{xy}(22)$	-0.22361	-0.21468	0.36631	0.00893
$\Sigma_{xy}(23)$	0.19365	0.19743	0.43212	0.00378
$\Sigma_{xy}(24)$	0.15811	0.17390	0.34203	0.01579
ρ_{xy}	0.30000	0.30405	0.07828	0.00405
$\Sigma 1_{xy}(11)$	1.00416	2.21826	4.54869	1.21410
$\Sigma 1_{xy}(12)$	1.56525	4.25959	10.03225	2.69434
$\Sigma 1_{xy}(13)$	1.73925	4.30226	9.20887	2.56301
$\Sigma 1_{xy}(14)$	1.42009	3.71180	8.74069	2.29170
$\Sigma 1_{xy}(21)$	-0.27386	-0.67614	4.06503	0.40228
$\Sigma 1_{xy}(22)$	-0.22361	-0.68435	4.28020	0.46074
$\Sigma 1_{xy}(23)$	0.19365	0.63663	3.26688	0.44298
$\Sigma 1_{xy}(24)$	0.15811	0.63454	3.27403	0.47643
$\rho 1_{xy}$	0.30000	0.46671	0.35035	0.16671

Model II*

Model II allows a different time element correlation for each partition. That is, the X values have their own time correlation and the Y values have their own time correlation. The XY cross correlation value have their own separate time correlation. For example the time correlation for the X values at time one and the X values at time two may be .1, the Y values time one to time two correlation may be .3. The X at time one and Y at time two may have a .2 correlation.

$$\begin{pmatrix} \Psi_y \otimes \Sigma_y & \Psi_{yx} \otimes \Sigma_{yx} \\ \Psi_{xy} \otimes \Sigma_{xy} & \Psi_x \otimes \Sigma_x \end{pmatrix} \tag{26}$$

Model III*

The model can retain the correlation structure in the time component of the Y values or the X values but not both. The cross correlation time structure retains its time correlation structure.

$$\begin{pmatrix} \Psi_y \otimes \Sigma_y & \Psi_{yx} \otimes \Sigma_{yx} \\ \Psi_{xy} \otimes \Sigma_{xy} & I_x \otimes \Sigma_x \end{pmatrix} \tag{27}$$

or

$$\begin{pmatrix} I_y \otimes \Sigma_y & \Psi_{yx} \otimes \Sigma_{yx} \\ \Psi_{xy} \otimes \Sigma_{xy} & \Psi_x \otimes \Sigma_x \end{pmatrix} \tag{28}$$

Either the Y or the X variable can be correlated in time. Hence the two equations 27 and 28 both have the same number of parameters to estimate.

$$\text{Parameters} = \frac{p(p+1)}{2} + \frac{q(q+1)}{2} + pq + 2 \tag{29}$$

Model IV*

This equation assumes the time correlation component is constant across all data. The mechanisms that are occurring in time play the same role in the X values, the Y values, and the cross project of the two. The corresponding matrix is shown in equation 30.

$$\begin{pmatrix} \Psi_t \otimes \Sigma_y & \Psi_t \otimes \Sigma_{yx} \\ \Psi_t \otimes \Sigma_{xy} & \Psi_t \otimes \Sigma_x \end{pmatrix} \tag{30}$$

In this model, all ρ parameters are equal.

$$\rho_x = \rho_y = \rho_{xy} \tag{31}$$

$$\text{Parameters} = \frac{p(p+1)}{2} + \frac{q(q+1)}{2} + pq + 1 \tag{32}$$

Model V*

The time correlation matrix Ψ equals the identity matrix throughout all four partitions.

$$\begin{pmatrix} I_y \otimes \Sigma_y & I_{yx} \otimes \Sigma_{yx} \\ I_{xy} \otimes \Sigma_{xy} & I_x \otimes \Sigma_x \end{pmatrix} \tag{33}$$

Model 33 shows no covariance structure across time units. This model assumes that what happens in time unit 1 does not influence time unit 2 or later. This simple analysis may be what a researcher will attempt when first faced with multivariate time series data.
Model 33 has the least number of parameters that require estimation.

$$\text{Model 33 Parameters} = \frac{p(p+1)}{2} + \frac{q(q+1)}{2} + pq \tag{34}$$

Hypothesis Testing

- Model I : completely unstructured covariance
- Model II : Kronecker product covariance structure with DC-8 and C-130 each having a different time correlation
- Model III : Kronecker product covariance structure where DC-8 or C-130 has no time correlation
- Model IV : Kronecker product covariance where DC-8 and C-130 have the same time correlation (A true maximum likelihood solution exists under transformation)
- Model V : Kronecker product covariance where there is no time correlation (A true maximum likelihood solution exists under transformation)

Hypothesis Testing*

The log likelihood ratio test was used to determine if the model should be increased in complexity. Note that the data is not normally distributed but this model was used as an approximate test statistics.

$$f(Z, \mu, \Sigma) = \frac{1}{2\pi^{(p+q)t/2}|\Sigma|^{1/2}} exp \frac{-(\mathbf{Z}')\Sigma^{-1}(\mathbf{Z})}{2} \tag{35}$$

$$\mathbf{Z} = \begin{bmatrix} Y \\ X \end{bmatrix} \tag{36}$$

The log likelihood ratio test was used as a test statistic.

$$-2(log f(Z, 0, \Sigma_{H_o}) - log f(Z, 0, \Sigma_{H_a})) \sim \chi^2_{df} \tag{37}$$

Hypothesis Testing*

Model	Sample Size	I vs II	II vs III	III vs V	II vs IV	IV vs V
II	500	.115	1.000	1.000	.782	1.000
II	350	.097	1.000	1.000	.290	1.000
II	200	.147	1.000	.999	.011	1.000
II	100	*	1.000	.993	0	1.000
II	50	*	.975	.982	0	1.000
III	500	.137	.109	1.000	1.000	1.000
III	350	.114	.105	1.000	1.000	1.000
III	200	.157	.094	1.000	.994	1.000
III	100	*	.084	1.000	.340	.965
III	50	*	.063	.996	.005	.732

Table: Rejection rates for 1000 samples

Hypothesis Testing*

Model	Sample Size	I vs II	II vs III	III vs V	II vs IV	IV vs V
IV	500	.113	1.000	.917	0	1.000
IV	350	.145	1.000	.923	0	1.000
IV	200	.145	1.000	.928	0	1.000
IV	100	*	1000	.889	0	1.000
IV	50	*	1000	.775	0	1.000
V	500	.122	.099	.021	0	.032
V	350	.091	.089	.015	0	.018
V	200	.142	.095	.020	0	.053
V	100	*	.080	.015	0	.038
V	50	*	.051	.023	0	.027

Table: Rejection rates for 1000 samples

Bootstrapping*

The hypothesis tests above showed a higher rejection rate than the expected .05. This was most likely due to differences between the MLE and the estimates. Bootstrapping was used to create tests that give more accurate rejection probabilities. Parametric bootstrapping stimulations based on (Efron and Tibshirani(1993)) theory were used to create hypothesis tests. For each possible covariance structure, a set of 100 simulations of 100 bootstrap samples each were used.

Bootstrap samples*

Initial data set.

$$\begin{pmatrix} y_{111} & \cdots & y_{p11} & \cdots & y_{1t1} & \cdots & y_{pt1} & x_{111} & \cdots & x_{q11} & \cdots & x_{1t1} & \cdots & x_{qt1} \\ \vdots & & \vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ y_{11n} & \cdots & y_{p1n} & \cdots & y_{1tn} & \cdots & y_{ptn} & x_{11n} & \cdots & x_{q1n} & \cdots & x_{1tn} & \cdots & x_{qtn} \end{pmatrix}' \tag{38}$$

The likelihood ratio statistics was used as a test statistic for the bootstrap.

$$\lambda = -2(log(L(Y, X, \mu, \Sigma_{H_o})) - log(L(Y, X, \mu, \Sigma_{H_a}).)$$

Bootstrapping*

The initial data was used to generate a bootstrap estimate of the variance covariance matrix

$$\Phi = \begin{pmatrix} \hat{\Psi}_y \otimes \hat{\Sigma}_y & \hat{\Psi}_{yx} \otimes \hat{\Sigma}_{yx} \\ \hat{\Psi}_{xy} \otimes \hat{\Sigma}_{xy} & \hat{\Psi}_x \otimes \hat{\Sigma}_x \end{pmatrix}.$$

The bootstrap variance covariance matrix was used to generate B bootstrap samples.

$$y_1^*, y_2^*, \ldots, y_b^*$$
$$x_1^*, x_2^*, \ldots, x_b^*$$

Bootstrapping*

Each bootstrap sample was used to create an estimate of the sample variance covariance matrix.

$$\Phi^* = \begin{pmatrix} \hat{\Psi}^*_y \otimes \hat{\Sigma}^*_y & \hat{\Psi}^*_{yx} \otimes \hat{\Sigma}^*_{yx} \\ \hat{\Psi}^*_{xy} \otimes \hat{\Sigma}^*_{xy} & \hat{\Psi}^*_x \otimes \hat{\Sigma}^*_x \end{pmatrix}.$$

Each bootstrap sample was also used to get a test statistics.

$$\lambda^*_b = -2(log(L(Y, X, \mu, \Sigma_{H^*_0})) - log(L(Y, X, \mu, \Sigma_{H^*_{(1)}}))).$$

λ was compared to the vector of λ^*_bs to get an estimate of the p-value.

Bootstrapping Hypothesis Tests*

Model	Sample Size	I vs II	II vs III	III vs V	II vs IV	IV vs V
II	500	.08	1.0	1.0	1.0	1.00
II	350	.07	1.0	1.0	.99	1.00
II	200	.06	1.0	1.0	.98	1.00
II	100	*	.92	1.0	.92	1.00
II	50	*	.95	.96	.66	1.00
III	500	.06	.05	1.00	1.00	1.00
III	350	.1	.06	1.00	1.00	1.00
III	200	.05	.01	1.00	1.00	1.00
III	100	*	.06	1.00	1.00	.96
III	50	*	.02	1.00	1.00	.77

Table: Rejection rates for 100 samples, Null Hypothesis II

Bootstrapping Hypothesis Tests*

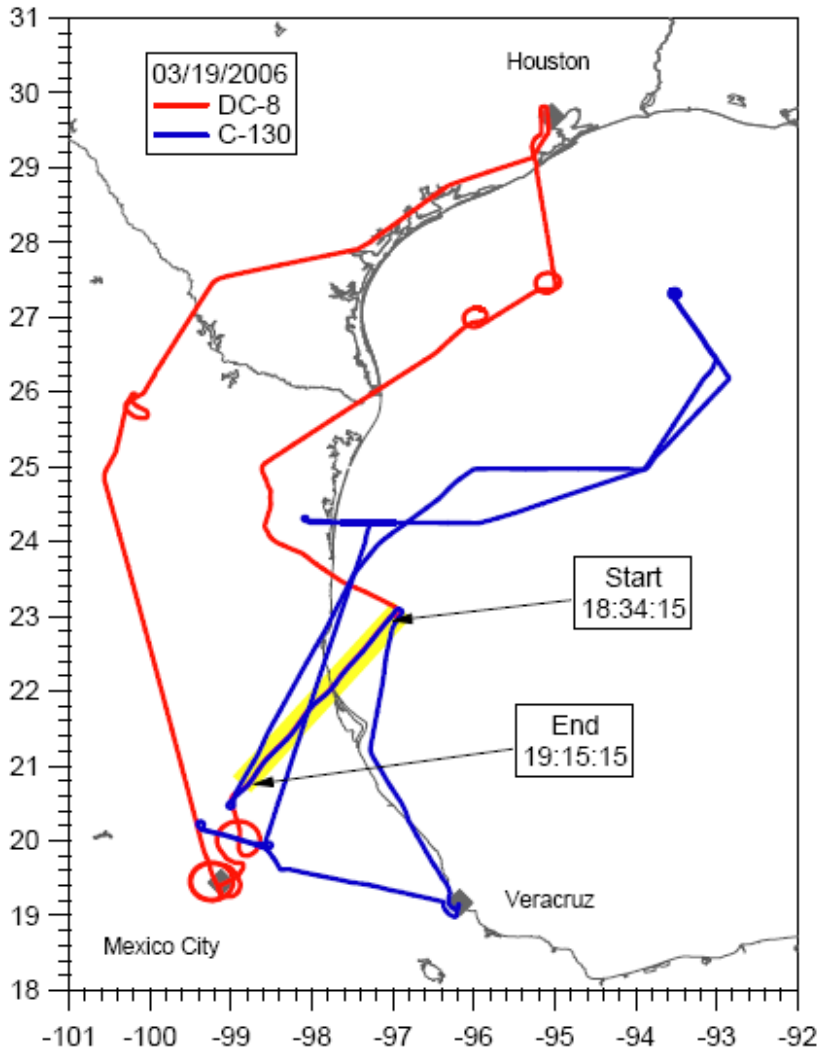
Model	Sample Size	I vs II	II vs III	III vs V	II vs IV	IV vs V
IV	500	.05	1.00	.94	.04	1.0
IV	350	.1	1.00	.89	.07	1.0
IV	200	.03	1.00	.94	.06	1.0
IV	100	*	1.00	.88	.06	1.0
IV	50	*	1.00	.88	.04	1.0
V	500	.05	.08	.1	.1	.07
V	350	.10	.08	.03	.07	.07
V	200	.03	.08	.09	.11	.06
V	100	*	.05	.08	.1	.04
V	50	*	.05	.05	.06	.06

Table: Rejection rates for 100 samples, Null Hypothesis II

INTEX*

- The Intercontinental Chemical Transport Experiment (INTEX)
- "INTEX (http://cloud1.arc.nasa.gov) is a two phase experiment that aims to understand the transport and transformation of gases and aerosols on transcontinental/intercontinental scales and assess their impact on air quality and climate."
- The experiment was performed in the spring of 2006.
- The purpose of the project was to "Quantify the outflow and evolution of gases and aerosols from the Mexico City Megaplex".

Analysis Air Tracks*



INTEX*

- Multiple air frames will measure air and pollutants along the Mexican Coast.
- NASA DC-8 flown out o Houston, Texas
- NSF/NCAR C-130 from Tampico, Mexico
- Air frames will travel in close proximity.
- Data from multiple gasses will be recorded for each plane and compared in an effort to calibrate the instrumentation.

INTEX*

- Three pollutants of interest for this study were H2O Water CO Carbon Monoxide O3 Ozone.
- Data were recorded over time and the three gasses were thought to be correlated.
- Sensors give different readings and neither sensor is considered to be the "correct" answer.
- The objective is to study the covariance structure of sensor measurement on both airframes. The structure will reveal how each plane's sensor readings vary with time.

Altitude and Molecule Measurements*

Altitude	O3 DC-8	CO DC-8	H20 DC-8	O3 C-130	CO C-130	H2O C-130
313	35.0317	112.93	2.295522603	33.2	103.3845	3.90521
3992.3	81.85337	222.59	14.55818318	85.2	211.3836	16.5065

Table: Altitude and Molecule Measurements before scaling

Units for CO and O3 are ppbv: Number of molecules per cubic centimeter over number of air molecules per cubic centimeter. Units for H20 are g/kg: grams of water vapor per kg dry air.

Altitude and Molecule Measurements*

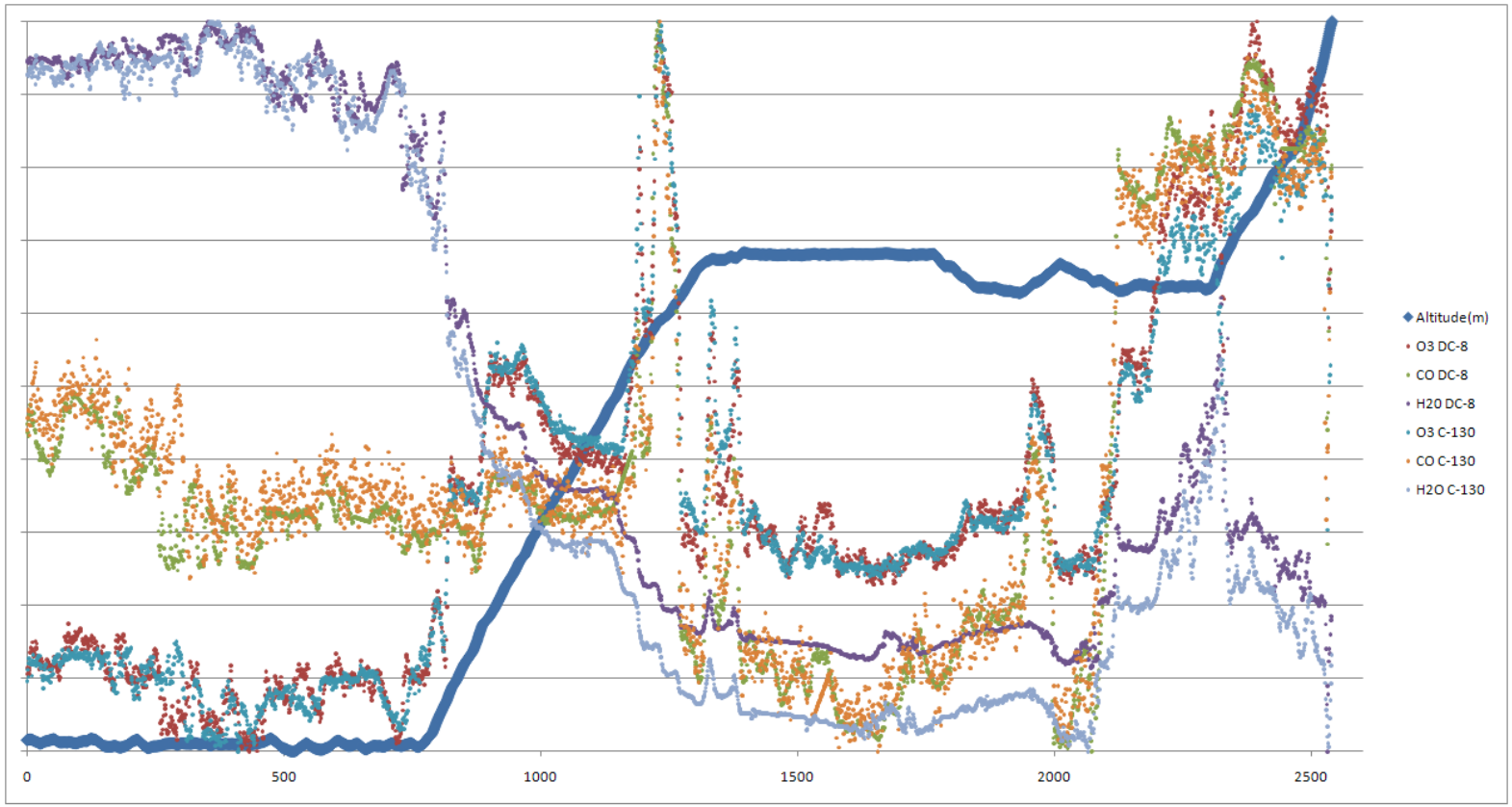


Figure: INTEX-B Airtracks Altitude

Data Hypothesis Test Results*

Hypothesis Test	Observed P-value
II $\Longrightarrow$ III	.00
II $\Longrightarrow$ IV	.23
IV $\Longrightarrow$ V	.007

Table: Testing Results for INTEX data

Variance Covariance Matrices*

Σ_yDC-8 O3	Σ_yDC-8 CO	Σ_yDC-8 H20
0.2119047	-0.011894	0.0049709
-0.011894	0.5828666	-0.01056
0.0049709	-0.01056	0.0053191

Table: Covariance IV Estimates for NASA Σ_y

Σ_xC-130 O3	Σ_x C-130 CO	Σ_xC-130 H20
0.2359842	0.1215785	0.0014622
0.1215785	7.2159748	0.0053734
0.0014622	0.0053734	0.011764

Table: Covariance IV Estimates for NASA Σ_x

Variance Covariance Matrices*

Σ_{xy} O3	Σ_{xy} CO	Σ_{xy} H20
0.0371183	-0.056899	-0.00096
0.0284117	-0.0285	0.002261
-0.000511	0.0094509	6.8809E-6

Table: Covariance IV Estimates for NASA Σ_{xy}

Ψ O3	Ψ CO	Ψ H20
1	-0.131819	0.0173763
-0.131819	1	-0.131819
0.0173763	-0.131819	1

Table: Covariance IV Estimates for NASA Ψ

Canonical Correlations*

Canonical Correlation	Correlation	Cumulative Percentage
1*st* Canonical Correlation	0.201539	.726285
2*nd* Canonical Correlation	0.046524	.893943
3*rd* Canonical Correlation	0.02943	1

Table: Canonical Correlations Within Each Time Period

Canonical Coefficients*

	1st Variable	2nd Variable	3rd Variable
Ozone DC-8	0.9296	-0.1022	-0.3845
Carbon M. DC-8	0.3716	0.6197	0.7178
Water DC- 8	-0.2272	0.9193	-0.4032

Table: Standardized Canonical Coefficients DC-8

	1st Variable	2nd Variable	3rd Variable
Ozone C-130	0.9602	0.2816	-0.0899
Carbon M. C-130	-0.3806	0.8357	-0.4072
Water C-130	-0.0589	0.3999	0.9152

Table: Standardized Canonical Coefficients C-130

Conclusion*

- Repeated CCA is a method that allows the comparison of multiple random variables to each other.
- The procedure is distribution independent and estimates the the variance covariance.
- The number of variables required to estimate variance covariance matricies grows quickly.
- Modeling the data struction in accordance with subject matter expert knowledge reduces the data requirements.

Bakewell, D., and Wit, E., "Bartlett Correction for likelihood ratio test", *Unpublished report*, 3 pages.

Bradley, E. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

Chaganty, N. R., and Naik, D. N. (2002), "Analysis of Multivariate Longitudinal Data using Quasi-Least Squares", *Journal of Statistical Planning and Inference*, 103, 421-436.

Harville, D. A. (1973), *Matrix Alegebra From a Statistician's Perspective*, New York: Springer.

Hotelling, H. (1936), "Relations Between Two Sets of Variates", *Biometrika*, 28, 321-377.


Johnson, R. A., and Wichern, D. W., (2002), *Applied Multivariate Statistical Analysis*, New Jersey: Prentice-Hall.


Kettenring, J. R. (1971), "Canonical Analysis of Several Sets of Variables", *Biometrika*, 58, 433-451.


Khattree, R., and Naik, D. N. (2000), *Multivariate Data Reduction and Discrimination with SAS Software.*, North Carolina: Wiley-SAS.


Mardia, K. V., Kent, J. J.(1979), & Bibby, J. M. (1979), *Multivariate Analysis*, New York: Academic Press.


Naik, D. N., and Rao, S. (2001), "Analysis of Multivariate Repeated Measures Data with a Kronecker Product Structured Covariance Matrix", *Journal of Applied Statistics*, 28, 91-105.

Ravishanker, N. and Dey, D (2002), *A First Course in Linear Models Theory*, New York: Chapman & Hall/CRC.

Roy, A., and Khattree, R. (2005), "On implementation of a test for Kronecker product covariance structure for multivariate repeated measures data", *Statistical Methodology*, 2, 297-306.

Schott, J. R. (1997), *Matrix Analysis for Statistics*, New York: John Wiley & Sons, Inc.

Srivastava J. (2007), "Canonical Variate Analysis and Related Methods with Longitudinal Data", *Ph.D. thesis, Department of Statistics, Old Dominions University*.

Srivastava, M. S., Nahtman, T., and von Rosen, D. (2008), "Estimation in general multivariate linear models with

Kronecker product covariance structure", *Research Report, Swedish University of Agricultural Sciences*, 21 pages.

Srivastava, M.S., Nahtman, T., and von Rosen, D.(2008), "Models with a Kronecker Product Covariance Structure: Estimation and Testing", *Mathematical Methods of Statistics*, 17(4), 357-370.

Srivastava, J and Naik, D. N. (2008), "Canonical Correlation Analysis of Longitudinal Data", *Denver JSM 2008 Proceedings, Biometrics Section*, 563-568.

Tan, W. Y. (1973), "Multivariate Studentization and its Applications", *The Canadian Journal of Statistics*, V1, N2, 181-199.

Questions?